Supplementary Information for

Identifying Sex-Specific Sub-phenotypes in Alzheimer's Disease Progression Using Longitudinal Electronic Health Records

Weimin Meng[1], Qiang Yang[1], Jie Xu[1], Yu Huang[1], Cankun Wang[2], Qianqian Song[1], Anjun Ma[2], Lixin Song[3], Jiang Bian[1], Qin Ma[2], Rui Yin[1, *]
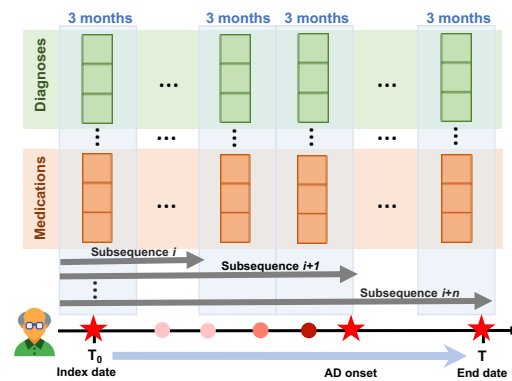
[1] Department of Health Outcomes & Biomedical Informatics, University of Florida, Gainesville, FL, USA
[2] Department of Biomedical Informatics, Ohio State University, Columbus, OH, 43210, USA
[3] School of Nursing, University of Texas Health Science Center at San Antonio, San Antonio, TX, 78229, USA
*Correspondence: ruiyin@ufl.edu (R.Y.)

Supplementary Figures

Supplementary Figure 1. The construction of Alzheimer's Disease (AD) temporal trajectory using EHRs. Relevant EHR data for each patient was aggregated in 3-month blocks (i.e., time intervals) into a set of vectors (e.g., Diagnoses, Medications). A vector contains unique clinical variables and their frequencies within a time interval. For the invariant variables, demographic, we treated them as static features and passed them at each time interval. We then divided each patient's vectors into multiple subsequences with varying time lengths, starting from the index date, and new subsequences are created with 3-month increments (i.e., 6-month, 9-month, etc.) until the last encounter of patients within the end date.

Supplementary Tables

Supplementary Table 1. Phecodes for the three most common comorbidities of AD.

Disease Category	Neurological disorders[1–3]	Cardiovascular diseases[4,5]	Diabetes[6,7]
Phecodes	Phe_Memory loss, Phe_Altered mental status, Phe_Transient alteration of awareness, Phe_Hallucinations, Phe_Schizophrenia and other psychotic disorders, Phe_Schizophrenia, Phe_Paranoid disorders, Phe_Mood disorders, Phe_Bipolar, Phe_Depression, Phe_Major depressive disorder, Phe_Suicidal ideation, Phe_Anxiety disorders, Phe_Anxiety disorder, Phe_Phobia, Phe_Personality disorders, Phe_Eating disorder, Phe_Anorexia nervosa, Phe_Sleep disorders, Phe_Insomnia.	Phe_Essential hypertension, Phe_Hypertension, Phe_Hypertensive heart and/or renal disease, Phe_Other forms of chronic heart disease, Phe_Ill-defined descriptions and complications of heart disease, Phe_Hypertensive heart disease, Phe_Congestive heart failure, nonhypertensive, Phe_Other chronic ischemic heart disease, unspecified, Phe_Heart failure with reduced EF [Systolic or combined heart failure], Phe_Acute rheumatic heart disease	Phe_Type 2 diabetes, Phe_Diabetes mellitus, Phe_Type 2 diabetes with renal manifestations, Phe_Type 2 diabetes with neurological manifestations, Phe_Polyneuropathy in diabetes, Phe_Type 1 diabetes, Phe_Type 2 diabetes with ophthalmic manifestations, Phe_Diabetes type 2 with peripheral circulatory disorders, Phe_Type 1 diabetes with neurological manifestations, Phe_Diabetes type 1 with peripheral circulatory disorders, Phe_Type 1 diabetes with ophthalmic manifestations, Phe_Type 2 diabetes with ketoacidosis

Supplementary Table 2. Descriptive statistics on the characteristics of the study cohort.

AD Cohort, N (%)	Total AD Patients (N = 1,665)	Female AD Patients (N = 961, 57.7%)	Male AD Patients (N = 704, 42.3%)
Demographics			
Age at AD diagnosis, mean (std)	76.7 (9.2)	77.1 (9.5)	76.1 (8.8)
Disease development, mean (std)			
Follow-up durations	8.1 (1.70)	8.21 (1.68)	7.95 (1.72)
Years before AD diagnosis	5.6 (1.72)	5.59 (1.73)	5.6 (1.7)
Years after AD diagnosis	2.5 (1.55)	2.62 (1.59)	2.36 (1.48)
Years from MCI to AD	2.26 (1.70)	2.26 (1.72)	2.27 (1.69)
Conversion rate from MCI to AD	266 (16.0%)	150 (15.6%)	116 (16.5%)
Years from AD to death	2.28 (1.45)	2.39 (1.48)	2.16 (1.42)
Age at death	81.7 (8.9)	82.1 (9.4)	81.2 (8.4)
Mortality rate, N (%)	209 (12.55%)	109 (11.34%)	100 (14.20%)
Hispanic, N (%)			
Hispanic	426 (25.6%)	248 (25.8%)	178 (18.5%)
Not Hispanic	1,230 (73.9%)	707 (73.6%)	523 (54.4%)
No Hispanic information	9 (0.5%)	6 (0.6%)	3 (0.3%)
Race, N (%)			
American Indian or Alaska Native	0 (0.0%)	0 (0.0%)	0 (0.0%)
Asian	10 (0.6%)	5 (0.5%)	5 (0.7%)
Black or African American	259 (15.6%)	157 (16.3%)	102 (14.5%)
White	1,298 (77.96%)	741 (77.11%)	557 (79.12%)
Multiple race	15 (0.90%)	6 (0.62%)	9 (1.28%)
Unknown	83 (5.0%)	52 (5.4%)	31 (4.40%)
Vital Signs			
Body Mass Index (BMI)	26.5 (5.11)	26.8 (4.21)	26.3 (5.71)
Smoking status, N (%)			
Non-smoker	125 (7.5%)	81 (8.4%)	44 (4.6%)
Current smoker	15 (0.9%)	10 (1.0%)	5 (0.5%)
Ex-smoker	87 (5.2%)	37 (3.9%)	50 (5.2%)
Unknown	1,438 (86.4%)	833 (86.7%)	605 (63.0%)

Comorbidity, N (%)			
Neurological disorders	1,483 (89.1%)	876 (91.2%)	607 (86.2%)
Cardiovascular diseases	696 (41.8%)	370 (38.5%)	326 (46.3%)
Diabetes	779 (46.8%)	439 (45.7%)	340 (48.3%)

Supplementary Table 3. Descriptive statistics on the characteristics of sex-specific AD progression subsequence clusters (i.e., states)

States, N (%) [out of N = 27,328]	C1 (N = 4,726, 17.3%)	C2 (N= 983, 3.6%)	C3 (N= 2,498, 9.1%)	C4 (N= 9,537, 34.9%)	C5 (N= 1,258, 4.6%)	C6 (N= 3,743, 13.7%)	C7 (N= 4,583, 16.8%)
Demographics							
Age at AD diagnosis, mean (std)	77.1 (9.27)	78.1 (9.85)	77.6 (9.54)	77.2 (9.43)	75.5 (9.27)	76.1 (8.73)	77.2 (8.35)
N (%) of female patients with a specific state, calculated among Female AD Patients (N = 961)	921 (95.8%)	554 (57.7%)	866 (90.1%)	846 (88.0%)	52 (5.4%)	3 (0.3%)	1 (0.1%)
N (%) of male patients with a specific state, calculated among Female AD Patients (N = 704)	704 (100.0%)	18 (2.6%)	4 (0.6%)	0 (0.0%)	514 (73.0%)	681 (96.7%)	531 (75.4%)
Disease development, mean (std)							
Follow-up durations	8.1 (1.7)	8.16 (1.66)	8.23 (1.67)	8.31 (1.65)	7.98 (1.69)	7.98 (1.7)	8.17 (1.6)
Years before AD diagnosis	5.59 (1.71)	5.55 (1.72)	5.59 (1.74)	5.65 (1.74)	5.56 (1.64)	5.61 (1.7)	5.73 (1.71)
Years after AD diagnosis	2.51 (1.55)	2.61 (1.6)	2.64 (1.59)	2.66 (1.58)	2.42 (1.5)	2.37 (1.49)	2.45 (1.49)
Years from MCI to AD	2.26 (1.7)	2.14 (1.62)	2.26 (1.69)	2.29 (1.68)	2.24 (1.66)	2.26 (1.74)	2.29 (1.79)
Conversion rate from MCI to AD	45 (2.7%)	36 (2.2%)	85 (5.1%)	136 (8.2%)	36 (2.2%)	85 (5.1%)	93 (5.6%)
Years from AD to death	2.28 (1.45)	2.34 (1.47)	2.34 (1.44)	2.4 (1.5)	2.36 (1.52)	2.16 (1.43)	2.18 (1.48)
Age at death	81.7 (8.9)	82.2 (9.1)	82.1 (9.7)	82.4 (9.3)	80.3 (8.9)	81.2 (8.4)	81.5 (8.3)
Hispanic, N (%)							
Hispanic	426 (25.6%)	142 (8.5%)	225 (13.5%)	223 (13.4%)	141 (8.5%)	173 (10.4%)	132 (7.9%)
Not Hispanic	1230 (73.9%)	426 (25.6%)	639 (38.4%)	619 (37.2%)	423 (25.4%)	508 (30.5%)	398 (23.9%)
No Hispanic information	9 (0.5%)	4 (0.2%)	6 (0.4%)	4 (0.2%)	2 (0.1%)	3 (0.2%)	2 (0.1%)
Race, N (%)							
American Indian or Alaska Native	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Asian	10 (0.6%)	3 (0.2%)	4 (0.2%)	5 (0.3%)	5 (0.3%)	5 (0.3%	4 (0.2%)
Black or African American	259 (15.6%)	83 (5.0%)	139 (8.3%)	143 (8.6%)	83 (5.0%)	102 (6.1%)	86 (5.2%)
White	1,298 (77.96%)	448 (26.9%)	669 (40.2%)	645 (38.7%)	448 (26.9%)	539 (32.4%)	418 (25.1%)
Multiple race	15 (0.90%)	4 (0.2%)	6 (0.4%)	6 (0.4%)	9 (0.5%)	8 (0.5%)	5 (0.3%)
Unknown	83 (5.0%)	34 (2.0%)	52 (3.1%)	47 (2.8%)	21 (1.3%)	30 (1.8%)	19 (1.1%)
Vital Signs							
Body Mass Index (BMI)	26.5 (5.1)	27.0 (4.3)	26.7 (4.1)	26.7 (4.1)	26.3 (5.7)	26.4 (5.9)	26.4 (5.9)

Smoking status, N (%)							
Non-smoker	125 (7.5%)	58 (6.0%)	76 (7.9%)	65 (6.8%)	36 (3.7%)	42 (4.4%)	36 (3.7%)
Current smoker	15 (0.9%)	7 (0.7%)	9 (0.9%)	7 (0.7%)	5 (0.5%)	5 (0.5%)	3 (0.3%)
Ex-smoker	87 (5.2%)	24 (2.5%)	36 (3.7%)	33 (3.4%)	45 (4.7%)	46 (4.8%)	34 (3.5%)
Unknown	1,438 (86.4%)	483 (50.3%)	749 (77.9%)	741 (77.1%)	480 (49.9%)	591 (61.5%)	459 (47.8%)
Comorbidity, N (%)							
Neurological disorders	1,483 (89.1%)	226 (39.5%)	521 (59.9%)	780 (92.2%)	219 (38.7%)	439 (64.2%)	470 (88.3%)
Cardiovascular diseases	696 (41.8%)	64 (11.2%)	191 (22.0%)	343 (40.5%)	105 (18.6%)	235 (34.4%)	276 (51.9%)
Diabetes	779 (46.8%)	133 (23.3%)	295 (33.9%)	403 (47.6%)	174 (30.7%)	273 (39.9%)	268 (50.4%)

Supplementary Table 4. Descriptive statistics on the characteristics of sex-specific AD progression sub-phenotypes.

Sub-phenotypes, N (%) [out of N = 1,403]	S1 [C1→C2→C3→C4] (N = 413, 24.8%)	S2 [C1→C3→C4] (N= 329, 19.8%)	S3 [C1→C5→C6→C7] (N= 346, 20.8%)	S4 [C1→C6→C7] (N= 178, 10.7%)	S5 [C1→C5→C6] (N= 137, 8.2%)
Demographics					
Age at AD diagnosis, mean (std)	77.2 (9.8)	76.9 (9.2)	76.2 (8.7)	77.1 (8.3)	74.8 (9.5)
Female, N (%)	406 (98.3%)	326 (99.1%)	1 (0.3%)	6 (3.4%)	2 (1.5%)
Male, N (%)	7 (1.7%)	3 (0.9%)	345 (99.7%)	172 (96.6%)	135 (98.5%)
Disease development, mean (std)					
Follow-up durations	8.3 (1.61)	8.29 (1.7)	8.24 (1.57)	8.04 (1.67)	7.39 (1.8)
Years before AD diagnosis	5.67 (1.73)	5.69 (1.76)	5.76 (1.66)	5.68 (1.8)	5.24 (1.52)
Years after AD diagnosis	2.64 (1.56)	2.6 (1.54)	2.48 (1.46)	2.37 (1.52)	2.14 (1.42)
Years from MCI to AD	2.25(1.65)	2.23(1.79)	2.29(1.8)	2.3(1.8)	2.23(1.68)
Conversion rate from MCI to AD	66 (15.9%)	53 (16.2%)	58 (16.7%)	29 (16.1%)	21 (15.3%)
Years from AD to death	2.27 (1.46)	2.31 (1.32)	2.26 (1.48)	1.93 (1.51)	2.37 (1.22)
Age at death	82.9 (9.0)	82.1 (10.1)	80.7 (8.8)	82.7 (7.1)	79.6 (8.7)
Hispanic, N (%)					
Hispanic	111 (26.9%)	84 (25.5%)	85 (24.6%)	43 (24.2%)	38 (27.7%)
Not Hispanic	293 (70.9%)	240 (72.9%)	260 (75.1%)	128 (71.9%)	97 (70.8%)
No Hispanic information	2 (0.5%)	2 (0.6%)	1 (0.3%)	1 (0.6%)	1 (0.7%)
Race, N (%)					
American Indian or Alaska Native	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Asian	3 (0.7%)	0 (0.0%)	3 (0.9%)	1 (0.6%)	1 (0.7%)
Black or African American	56 (13.6%)	60 (18.2%)	55 (15.9%)	26 (14.6%)	12 (8.8%)

White	321 (77.7%)	245 (74.5%)	279 (80.6%)	133 (74.7%)	109 (79.6%)
Multiple race	2 (0.5%)	2 (0.6%)	3 (0.9%)	2 (1.1%)	3 (2.2%)
Unknown	124 (30.0%)	89 (27.1%)	122 (35.3%)	56 (31.5%)	37 (27.0%)
Vital Signs					
Height, inch	62.7 (4.8)	67.1 (57.9)	69.0 (3.1)	69.0 (2.9)	70.8 (17.6)
Weight, lbs.	152.8 (32.6)	154.3 (34.4)	193.0 (40.6)	189.1 (33.8)	182.2 (36.4)
Diastolic Blood Pressure, mmHg	72.6 (11.3)	73.3 (12.2)	76.2 (11.6)	75.7 (12.3)	74.8 (10.5)
Systolic Blood Pressure, mmHg	135.8 (21.6)	135.4 (21.4)	134.6 (20.4)	133.2 (17.6)	133.6 (22.3)
Body Mass Index (BMI)	26.9 (5.1)	27.0 (5.8)	27.9 (4.3)	27.7 (4.6)	27.9 (4.3)
Smoking status, N (%)					
Non-smoker	26 (6.3%)	11 (3.3%)	14 (4.0%)	5 (2.8%)	2 (1.5%)
Current smoker	0 (0.0%)	0 (0.0%)	3 (0.9%)	1 (0.6%)	0 (0.0%)
Ex-smoker	6 (1.5%)	6 (1.8%)	17 (4.9%)	7 (3.9%)	5 (3.6%)
Unknown	144 (34.9%)	109 (33.1%)	152 (43.9%)	66 (37.1%)	37 (27.0%)
Comorbidity, N (%)					
Neurological disorders	149 (36.1%)	113 (34.3%)	143 (41.3%)	70 (39.3%)	37 (27.0%)
Cardiovascular diseases	68 (16.5%)	65 (19.8%)	78 (22.5%)	34 (19.1%)	21 (15.3%)
Diabetes	171 (41.4%)	170 (51.7%)	178 (51.4%)	85 (47.8%)	50 (36.5%)

Supplementary References

1. Lyketsos, C. G. & Olin, J. Depression in Alzheimer's disease: overview and treatment. *Biol. Psychiatry* **52**, 243–252 (2002).
2. Ju, Y.-E. S. *et al.* Sleep quality and preclinical Alzheimer disease. *JAMA Neurol.* **70**, 587–593 (2013).
3. Wang, C. & Holtzman, D. M. Bidirectional relationship between sleep and Alzheimer's disease: role of amyloid, tau, and other factors. *Neuropsychopharmacology* **45**, 104–120 (2020).
4. Newman, A. B. *et al.* Dementia and Alzheimer's disease incidence in relationship to cardiovascular disease in the Cardiovascular Health Study cohort. *J. Am. Geriatr. Soc.* **53**, 1101–1107 (2005).
5. Rosenberg, P. B. *et al.* Effects of cardiovascular medications on rate of functional decline in Alzheimer disease. *Am. J. Geriatr. Psychiatry* **16**, 883–892 (2008).
6. Wang, X.-F. *et al.* Linking Alzheimer's disease and type 2 diabetes: Novel shared susceptibility genes detected by cFDR approach. *J. Neurol. Sci.* **380**, 262–272 (2017).
7. Janson, J. *et al.* Increased risk of type 2 diabetes in Alzheimer disease. *Diabetes* **53**, 474–481 (2004).